The Journal of Transport and Land Use http://jtlu.org
Vol. 11 No. 1 [2018] pp. 49–65

JTLU

A multi-dimensional multi-level approach to measuring the spatial structure of U.S. metropolitan areas

Arefeh Nasri
University of Maryland
aanasri@umd.edu

Lei Zhang (corresponding author)
University of Maryland
lei@umd.edu

Abstract: For many years, attempts to measure the urban structure and physical form of metropolitan areas have been focused on a limited set of attributes, mostly density and density gradients. However, the complex nature of the urban form requires the consideration of many other dimensions to provide a comprehensive measure that includes all aspects of the urban structure and growth pattern at different hierarchical levels. In this paper, a multi-dimensional method of measuring urban form and development patterns in urban areas of the United States is presented. The methodology presented here develops several variables and indices that contribute to the characterization and quantification of the overall physical form of urban areas at various hierarchical levels.

Cluster analysis is performed to group metropolitan areas based on their urban form and land-use pattern. This allows for a better utilization of land-use transportation planning and policy analyses used by planners and researchers. This clustering of urban areas could eventually help policymakers and decision makers in the decision-making process to evaluate land-use transportation policies, identify similar patterns, and understand how similar policies implemented in urban areas with similar urban form structure would result in more efficient and successful planning in the future.

Keywords: Built environment, spatial analysis, land-use, land use, metropolitan structure, sprawl, cluster analysis, urban form.

Article history:
Received: February 22, 2017
Received in revised form: September 16, 2017
Accepted: September 30, 2017
Available online: January 5, 2018

1 Introduction

There is an immense body of literature on the definition, quantification, and analysis of urban sprawl (see Ewing, 1994, 1997; Frenkel & Ashkenazi, 2008; Galster et al., 2001; Malpezzi & Guo, 2001; Torrens & Alberti, 2000; McCann & Ewing, 2003). However, sprawl is only part of the overall spatial structure of urban areas. Research on quantifying urban structure at the metropolitan scale is very limited in the body of literature both from theoretical and empirical points of view.

For example, Tsai (2005) categorized the overall metropolitan form into three types using four dimensions of size, density, degree of equal distribution, and degree of clustering. He also used a Moran

http://dx.doi.org/10.5198/jtlu.2018.893
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 3.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

coefficient to distinguish among monocentric (high values), polycentric (intermediate values), and decentralized (low values) urban forms. Later on, Veneri (2010) used measures such as a polycentricity index, spatial concentration of employment, gross density, mixed land use, transit coverage, metropolitan size (area), and population structure index (age distribution) to quantify the metropolitan-wide urban form. More recently, Yang, French, Holt, and Zhang (2012) developed new urban form metrics such as the spatial variation of density, the density of suburban centers relative to the region, and the spatial distribution of high-density nodes, and examined the impacts of their metropolitan-wide measures on commuting time in the 50 largest U.S. metropolitan areas from 1970 to 2000.

The number of studies focusing on urban form at the macro and regional scales and its inter-relationship with travel behavior is also relatively limited (Gordon, Kumar, & Richardson, 1989; Gordon, Richardson, & Jun, 1991; Grengs, Levine, Shen, & Shen, 2010; Ingram, 1998; Nasri & Zhang, 2012, 2014). For instance, Bento et al. (2005) examined the effect of metropolitan urban form on commute mode choice and vehicle miles traveled (VMT) using data from the 1990 Nationwide Personal Transportation Survey. They developed measures such as population centrality (the spatial distribution of population — how close to the city center the population is located), jobs-housing balance, city shape (how close to circular the city is), and road density. Another study by Kelly-Schwartz, Stockard, Doyle, and Schlossberg (2004) investigated the effect of urban form on health by developing measures of density, centrality (strength of metropolitan centers), accessibility (street network density and interconnectivity), and neighborhood mix (mix of work, shopping, and housing). Cervero and Murakami (2010) performed an analysis of the effect of urban form on VMT using a sample of 370 urban areas across the United States. They used measures such as population and road network densities as well as job and retail accessibility, and found a strong causal relationship between the overall urban structure of metropolitan areas and per capita VMT. However, as most of these studies pointed out, their analysis does not include a complete and comprehensive measure of urban form due to data limitation and the fact that there is no consensus on what a complete measure is and how it would be calculated with the current data availability.

In summary, the characterization/quantification of the overall spatial form of metropolitan areas is still a challenge, requiring further steps than only measuring the degree of sprawl to fully address this big picture. Thus, our focus in the present study is how to measure the overall form of urban areas and large-scale built environment pattern using a wide range of variables. The present study significantly contributes to the literature as it tries to overcome most of the data limitations faced by previous studies by using a disaggregated and very detailed land-use data at the national level covering a wide range of different aspects of urban form. It is one of the first attempts to create a multi-dimensional picture of the overall form of urban areas, including measures of sprawl as well as other spatial measures to quantify the built environment. It shows that the overall spatial structure of urban areas is not just measured by the degree of sprawl, but also by other factors such as the degree of centrality and accessibility, housing composition and urban morphology, and network structure, job diversity and concentration. The results will eventually help regional planning agencies and decision makers in their efforts to investigate interrelationships between urban form at the metropolitan scale and individual travel pattern and lifestyle.

We measure variables that can only be measured at the metropolitan level and variables that are usually measured at a smaller scale, aggregating them to higher levels, since those variables convey different meanings when measured at different scales (Tsai, 2005). This approach is thus called a multilevel approach as for several variables, calculation was first done at the census block group (CBG) level (which is finest level at which the data was available) and then aggregated to the metropolitan level. A comprehensive quantification of urban form provided by various measures allows for a better understanding and visualization of various aspects of urban form; it can also facilitate and enrich the analysis

of the relationships among urban structure and various dimensions of urban daily travel behavior (i.e., VMT, travel time, trip distance, mode choice, departure time choice, and destination choice) as well as the planning and evaluation of various land-use policy scenarios, such as transit-oriented development, smart growth, and polycentricity. It can also utilize research in areas other than travel behavior analysis, such as environmental analysis, housing markets, and economic development.

Measures of urban form introduced in this research belong to two separate groups; those measured at the local and neighborhood levels and then aggregated/averaged to represent the macro-scale characteristics and those directly measured for the region/metro area as a whole. The top 50 metropolitan areas in the United States in terms of 2010 population, according to U.S. Census 2010, are selected as case study areas, and the proposed measures and indices have been calculated for these urban areas. The consistency of the values allows for several comparative analyses in the 50 metropolitan areas of study.

Table 1 lists all these metropolitan areas along with their 2010 population and employment. Among these cases, New York is the first rank in terms of both population (18,897,109) and employment (8,022,279); Birmingham, AL, and Salt Lake City, UT, have the lowest employment (477,549) and population (1,124,197), respectively. In terms of geometric area, Riverside, CA, (17,548,869.82 acre) is the largest and Hartford, CT, (1,028,311.86 acre) is the smallest metro area among all.

Although the metropolitan areas of study all share high population and employment, they are not similar in every characteristic — especially in terms of their urban form and built environment pattern. They vary in size (i.e., developable land area), densities, accessibilities, housing characteristics, road network structure, and more. This paper tries to address these differences and help finding patterns among urban form, travel behavior, and transportation system performance (e.g., level of congestion and transit ridership rates). To achieve this goal, the cluster analysis method was used to investigate the similarities and differences among the cities in terms of their urban structure and transportation supply patterns. Cases have been grouped based on their spatial and urban form characteristics into three categories of 1) compact, well-mixed, high-accessible, 2) moderate-density, reasonable accessibility and connectivity pattern, and 3) sprawled, low-density, suburban setting. This classification could help urban planners and policy makers for policy analysis and decision-making process, based on comparative analyses that result in similar cities implementing different planning/policy strategies.

Table 1: Case study areas

Metropolitan area	Population	Employment	Metropolitan area	Population	Employment
Atlanta, GA	5,268,860	2,203,331	Minneapolis-St. Paul, MN	3,279,833	1,679,161
Austin, TX	1,716,289	800,514	Nashville, TN	1,589,934	742,661
Baltimore-Towson, MD	2,710,489	1,212,756	New Orleans, LA	1,167,764	495,052
Birmingham-Hoover, AL	1,128,047	477,549	New York, NY-NJ	18,897,109	8,022,279
Boston-Cambridge, MA	4,552,402	2,338,890	Oklahoma City, OK	1,252,987	546,958
Buffalo-Niagara Falls, NY	1,135,509	542,353	Orlando, FL	2,134,411	978,967
Charlotte, NC	1,758,038	770,971	Philadelphia, PA	5,965,343	260,046
Chicago, IL	9,461,105	4,161,510	Phoenix, AZ	4,192,887	1,661,476
Cincinnati, OH	2,130,151	944,787	Pittsburgh, PA	2,356,285	1,093,445
Cleveland, OH	2,077,240	957,557	Portland, OR	2,226,009	974,858
Columbus, OH	1,836,536	865,988	Providence, RI	1,600,852	661,822
Dallas, TX	6,371,773	2,871,213	Raleigh-Cary, NC	1,130,490	548,185
Denver, CO	2,543,482	1,212,658	Richmond, VA	1,258,251	571,928
Detroit, MI	4,296,250	1,657,054	Riverside, CA	4,224,851	1,183,673
Hartford, CT	1,212,381	599,586	Sacramento, CA	2,149,127	840,310
Houston, TX	5,946,800	2,530,059	Salt Lake City, UT	1,124,197	592,557
Indianapolis-Carmel, IN	1,756,241	864,558	San Antonio, TX	2,142,508	801,317
Jacksonville, FL	1,345,596	653,161	San Diego, CA	3,095,313	1,230,279
Kansas City, MO-KS	2,035,334	941,315	San Francisco, CA	4,335,391	1,953,826
Las Vegas-Paradise, NV	1,951,269	806,758	San Jose, CA	1,836,911	866,354
Los Angeles, CA	12,828,837	5,566,994	Seattle, WA	3,439,809	1,600,098
Louisville/Jefferson County, KY-IN	1,283,566	586,897	St. Louis, MO-IL	2,812,896	1,261,547
Memphis, TN-MS-AR	1,316,100	570,014	Tampa, FL	2,783,243	1,046,561
Miami, FL	5,564,635	2,118,833	Virginia Beach-Norfolk, VA-NC	1,671,683	674,996
Milwaukee, WI	1,555,908	794,235	Washington, DC	5,582,170	2,781,078

2 Data, variables, and calculation process

Georeferenced land-use data was mainly obtained from the Smart Location Database (SLD), which was developed by the Environmental Protection Agency (EPA). It provides demographics, employment, and built environment measures at the census block group level. This rich nationwide dataset develops the five Ds including residential and employment *density*, land-use *diversity*, *design* of the built environment, *access to destinations*, and *distance* to transit as the main built environment characteristics using several different data sources such as Census TIGER/line data, Census LEHD[1], NAVSTREETS[2], etc. SLD and the corresponding GIS shapefiles have been used to conduct a spatial analysis of the urban form and calculate the built environment variables.

More than 50 variables were developed for this study and are listed in Table 2 under five categories of socioeconomic and demographic, housing and urban morphology, density and centrality, diversity and urban design, and network and destination accessibility.

[1] Longitudinal Employer-Household Dynamics

[2] Provided by NAVmart, global Resource Center for geospatial data; https://navmart.com/here-navstreets/

Table 2: Variable description and data sources

Variables	Description	Data source
Socioeconomic and demographics		
EmpTot	Employment, 2010	SLD
PopTot	Population, 2010	SLD
HHs	Number of households (occupied housing units), 2010	SLD
Workers	# of workers (home location), 2010	SLD
Avg_HH_size	Average household size/aggregated from CBGs	SLD
P_WrkAge	Percent of working-age population, 2010	SLD
MedHHInc	2010 median household income in the CBSA[3]	HUD*
P_AutoOwn0	Percent households with zero cars in the CBSA	SLD
P_AutoOwn1	Percent households with one car in the CBSA	SLD
P_AutoOwn2+	Percent households with 2+ cars in the CBSA	SLD
P_LowWage	Percent workers earning $1,250/month or less (home location), 2010	Derived from SLD
P_MedWage	Percent workers earning more than $1,250/month but less than $3,333/month (home location), 2010	Derived from SLD
P_HiWage	Percent workers earning $3,333/month or more (home location), 2010	Derived from SLD
P_CrossCommuter	Percent of employment that commutes in/out of metro area	Derived from SLD
Housing and urban morphology		
HH_type1_h	Housing cost as % of income for a median-income family	LAI**
HH_type7_h	Housing cost as % of income for a moderate-income family	LAI
HH_type8_h	Housing cost as % of income for a high-income family	LAI
P_unprotected	Percent geometric area (acres) that is not protected from development (i.e., not a park or conservation area)	Derived from SLD
P_occupied	Percent of occupied housing units in the CBSA	Derived from SLD
Avg_Occupation	Average percent of occupied housing units (from CBGs)	Derived from SLD
Density and centrality		
ResDens_Avg	Average residential density	Derived from SLD
EmpDens_Avg	Average employment density	Derived from SLD
StDev_Popdens	Standard deviation of population density	Derived from SLD
StDev_Empdens	Standard deviation of employment density	Derived from SLD
CoV_Popdens	The coefficient of variation of population density	Derived from SLD
CoV_Empdens	The coefficient of variation of employment density	Derived from SLD
P_ResOnly	Percent population living in residential-only CBGs	Derived from SLD
P_LowResDens	Percent population living in low-residential-density zones	SLD- Spatial analysis
P_Hi_ResDens	Percent population living in high-residential-density zones	SLD-Spatial analysis
P_ LowEmpDens	Percent population living in low-employment-density zones	SLD-Spatial analysis
P_ HiEmpDens	Percent population living in high-employment-density zones	SLD-Spatial analysis
E_LowEmpDens	Percent employment in low-employment-density zones	SLD-Spatial analysis
E_HiEmpDens	Percent employment in high-employment-density zones	SLD-Spatial analysis
Diversity and urban design		
Entropy_Avg	CBG land-use mix score/averaged for metropolitan area	Derived from SLD
Job_HH_Avg	Jobs per HH at CBG level/averaged for metropolitan area	Derived from SLD
%SmallBlocks	Percent blocks smaller than 0.01 sq. mi	Census/TIGER 2010
Block_Size_Avg	Average block size/Aggregated from CBGs	Census/TIGER 2010

[3] Core-based statistical area

Table 2: Variable description and data sources *(cont.)*

Variables	Description	Data source
Network and destination accessibility		
Rd_metro	Total road network density	Census/TIGER 2010
IntDens_metro	Total intersection density	Derived from SLD
P_Trans_Pop	Percent population living within ½ mile of transit	Derived from SLD
P_Trans_Emp	Percent jobs located within ½ mile of transit stops	Derived from SLD
PJ_45_auto	Percent jobs within 45 minutes auto travel time	Derived from SLD
PJ_45_transit	Percent jobs within 45-minute transit travel time	Derived from SLD
PW_45_auto	Percent working age population within 45 minutes auto travel time	Derived from SLD
PW_45_transit	Percent working-age population within 45-minute transit travel time	Derived from SLD
RetAcc_avg	Average ratio of residential population to retail employment	Derived from SLD
P_NoRet	Percent population living in no-retail zones	Derived from SLD
WalkScore	Walk score/ walkability at the metropolitan level	WalkScore Inc.
Congestion_Index	Level of congestion in a metro area	TTI***

* U.S. Department of Housing and Urban Development

** Location Affordability Index Data

*** Texas Transportation Institute

Most of the socioeconomic and demographic variables listed in the table above, are available at the CBG level in the SLD database and have been used to obtain the aggregated values at the metropolitan level. The percent of workers in different income group (low-, median-, and high- wage groups) is calculated by summing up the number of workers in each group and dividing that value by the total number of workers in the metropolitan area. The percentage of cross-commuters is calculated by subtracting the number of workers from the total employment in a metropolitan area. If that number is positive, it implies that workers from the outside region have to commute and fill in the excess employment opportunities (commuters-in). Similarly, if the number is negative and the number of workers is greater than the total employment, the excess workers have to commute to outside region for work (commuter-out). If the number of workers and the total employment in the metro area are equal, there are no cross-commuters and the value for this variable is equal to zero.

In terms of urban form measures, in addition to the metropolitan-wide population and employment densities that provide valuable dimensions of urban structure (Schwanen, Dieleman, & Dijst, 2004), it is important to identify the spatial variation of population and employment densities, as well as the spatial distribution of high-density nodes and the spatial clustering pattern and the degree of such clustering —whether it is clustering of low or high values — as all these factors appear to influence travel pattern in urban areas distinctively (Yang et al., 2012).

Spatial analysis was done in ArcGIS to calculate several variables under the density and centrality category. For each of the case study areas, the spatial statistics tool in ArcMap 10.1 was used to investigate the existence of high/low population and employment density clusters and then to identify the location and distribution of such clusters.

The spatial autocorrelation (Global Moran's I) tool assesses the overall pattern and trend of the data used. They are most effective when the spatial pattern is consistent across the study area, whereas the local statistics (like the hot-spot analysis tool) assess each feature within the context of neighboring features, comparing the local situation to the global situation. Similarly, global spatial statistics, including the spatial autocorrelation (Global Moran's I) tool, are not effective when the variable being measured is not consistent across the entire study area. As a result, the high/low clustering (Getis-Ord general G)

tool is most appropriate when we are looking for unexpected spatial spikes of high/low values. It identifies the concentration of high and low values of a certain feature and computes a z-score describing the degree of spatial concentration or dispersion for a certain variable (Fischer & Getis, 2009). The high/low clustering (Getis-Ord General G) tool is an inferential statistic, which means that the results of the analysis are interpreted within the context of the null hypothesis. The null hypothesis states that there is no spatial clustering of feature values. When the test is done and the p-value returned by this tool is small and statistically significant, the null hypothesis can be rejected. Once the null hypothesis is rejected, then we look at the sign of the z-score. The positive values for the z-score indicate the hot spots (clusters of high values) and the negative values for the z-score indicate the cold spots (clusters of low values) of a certain feature.

Using the spatial statistics tools in ArcGIS 10.1, the high/low clustering (Getis-Ord General G) test was performed to measures the degree of clustering for either high values or low values for the absolute population and employment values as well as for population and employment densities. It also identified the location and size of such clusters. The results show that in most cases, there is a clustering of low values for population and a clustering of high values for employment, which confirms the suburban setting for these cases where people live in low-density decentralized residential zones with employment opportunities concentrated in centers far from residential zones. This, consequently, is associated with an auto-oriented life style. Once those clusters are identified, the overall population and employment located within the high-density zones was calculated. The resulting six variables are: P_LowResDens, P_Hi_ResDens, P_ LowEmpDens, P_ HiEmpDens, E_LowEmpDens, and E_HiEmpDens (see Table 2 for variable descriptions).

The housing cost as a percentage of income has been obtained from the Location Affordability Index data (LAI) for three types of households; median-income family (based on the region's median income), moderate-income family (earning 80% of the region's median income), and high-income family (earning 150% of region's median income). This dataset provides housing and transportation cost as a percentage of households' income at multiple levels including state, county, city, census tract, and census block group levels. The data was downloaded at the CBG level and aggregated to obtain the average housing cost for each of the case study areas.

The congestion_index variable was obtained from the Texas Transportation Institute (TTI). They calculated this index for about 500 metropolitan areas nationwide using a variety of data sources on traffic volume, speed, and average travel time. The traffic speed was obtained from INRIX, a private company that provides travel time information for each section of road for every 15 minutes of each day, for a total of 672 day/time period cells (24 hours x 7 days x 4 periods per hour).

The rest of the variables have been either directly obtained from the SLD or calculated using the SLD and other data sources, such as Census TIGER shapefiles as listed in Table 2 above. Several of these variables were then used to conduct a cluster analysis to group the metropolitan areas with similar land-use pattern.

3 Cluster analysis

After the land-use measures are calculated for each of the case study area, it is observed that although the metropolitan areas of study all share high population and employment, they are not similar in every characteristic especially in terms of their urban form and built environment pattern. They vary in size (i.e., developable land area), densities, accessibilities, housing characteristics, road network structure, and more. Cluster analysis was performed to group the similar cities together based on their overall urban form pattern and investigate the similarities and differences between groups of cities in terms of their urban structure, transportation supply patterns, and aggregate-level travel behavior.

Cluster analysis has a wide range of applications in many research fields such as marketing, insurance, biology, and psychiatry (see Kaufman & Rousseeuw, 1990; Punj & Stewart, 1983; Borgen & Barnett, 1987). In land-use planning and policy-making, cluster analysis can be very useful to identify areas with similar land-use pattern to propose, implement, and evaluate land-use policies more efficiently (see Smith & Saito, 2001 as an example).

In our cluster analysis, Euclidean distance measure was used as a measure of similarity to form groups of observations (i.e., cities). It is calculated by:

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{ip} - x_{jp}|^2} \quad (1)$$

Clustering algorithms can be categorized into several groups, such as partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. In this analysis, we only describe the two groups of partitioning methods and hierarchical methods. Introducing and analyzing the rest of the methods is beyond the scope of this analysis.

In the partitioning approach, the database is partitioned into a set of n clusters with the minimum sum of squared distance. The process begins with n initial group centers; each observation is assigned to the cluster group to which its mean or median is the closest. The process is repeated until all the observations belong to the cluster group with the closest mean/median to the center and no observation changes group. There are two main methods in this approach; k-means and k-median. In the k-means method, each cluster is represented by the center of the cluster. The algorithm iteratively estimates the cluster group means and assigns each observation to the cluster for which its distance to the cluster mean is the smallest. The process is continued until no observation changes group. In the k-median method, each cluster is represented by one of the observations in the cluster (the most centrally located observation in a cluster).

In the hierarchical clustering approach, the distance matrix is used as clustering criteria. It is not required to specify the number of clusters in advance. Instead, it needs a termination condition. In this method, data is decomposed into several levels of nested partitioning which is called dendrogram. The clustering of observations is obtained by cutting the dendrogram at a desired level (based on the number of clusters needed). Hierarchical clustering is categorized -based on the distance measure used- into several methods such as single linkage, average linkage, complete linkage etc. (see Table 3 for a complete list and description of clustering methods).

Table 3: Cluster analysis methods summary and descriptions

Method	Description	# of Clusters
Partition clustering methods		
Kmeans	Each cluster is associated with a centroid/ Construct various partitions in which each observation belongs to the cluster with the nearest mean.	User-specified mandatory
Kmedians	A variation of kmeans clustering. The same process is followed except that medians, instead of means, are computed to represent the group centers at each step.	
Hierarchical clustering methods		
Single linkage (nearest-neighbor method)	The closest two groups are determined by the closest observations between the two groups.	Optional/ Done in post-clustering
Average linkage (arithmetic-average clustering)	The closest two groups are determined by the average (dis)similarity/ distance between the observations of the two groups.	
Complete linkage (furthest-neighbor method)	The closest two groups are determined by the farthest observations between the two groups.	
Weighted average linkage (weighted group-average method)	Similar to average-linkage clustering, except that it gives each group of observations equal weight, while average linkage gives each observation equal weight.	
Median linkage (weighted pair method)	A variation on centroid linkage; treats groups of unequal size differently. It gives each group of observations equal weight, meaning that with unequal group sizes, the observations in the smaller group will have more weight than the observations in the larger group.	
Centroid linkage (unweighted pair-group centroid method)	Merges the groups whose means are closest. Gives each observation equal weight.	
Ward's linkage (minimum-variance method)	Produces clusters of similar numbers of observations and with a minimal amount of within-cluster variance.	

Several cluster analysis methods—including k-means, average linkage, complete linkage, and Ward's linkage methods—and different number of clusters were applied and tested on the data. A combination of k-means method and the Euclidean distance measure using six main land-use variables was selected as it produced the most logical clustering of metropolitan areas of study (based on the output's similarity indices). The land-use variables based on which the final clustering was performed are as follows: average employment density in the metro area, average population density in the metro area, average entropy score, retail accessibility in the metro area, average block size, and proportion of metro area's employment within ½ mile of major transit stops (transit accessibility). The clustering process starts with several combinations of the urban form variables calculated in the previous step and the clustering results are compared using the cluster analysis evaluation and performance measures. The final clustering is performed with the six variables listed above and the others were dropped due to high ratio of collinearity or weak performance measures for the clustering output. The socio-demographic variables were not included in the clustering process as these variables are not part of the urban form and built environment characteristics of the cities. However, they are calculated for every city in our analysis, as these variables are very useful in understanding the population composition, travel behavior preferences, and residential location choice/preferences in urban areas.

Cities have been grouped based on several of their spatial and urban form characteristics into three categories of 1) compact, well-mixed, high-accessible, 2) moderate-density, average accessibility, random clustering pattern and 3) sprawled, low-density, suburban setting. This reasonable set of classifications could facilitate research on various aspects of land use and transportation interactions in different urban areas and serve as a guideline to help urban planners and policy makers better understand the relationships between the overall land-use pattern and travel outcomes in certain urban areas. It is also useful for the decision-making process and development and evaluation of various land-use policy scenarios based on comparative analyses results, especially considering the similarities/differences of cities in the same cluster groups.

Table 4 represents the three cluster types obtained using the k-means clustering method and lists the metropolitan areas falling under each of the three types. As it is indicated, Cluster type A (compact, well-mixed, high-accessible cities) consists of seven cases, most of which are among the top ten metropolitan areas in terms of the overall population and employment. However, it also shows that not necessarily all the cases with high population and employment have an overall dense and highly accessible urban structure. For example, Los Angeles and Dallas, which are among the top five metropolitan areas both in terms of population and employment, are not categorized under cluster type A. The cluster type B consists of 25 cases and is identified as a group of moderate-density cities with reasonably good job accessibility and street connectivity. Cases in this group range from Los Angeles and Dallas, with overall high population and employment to Buffalo and New Orleans, which are among the small low-population and employment cities. Finally, cluster type C with 18 cases is identified as the group of suburban style cities with overall sprawled low-density pattern and low job accessibility and walkability.

Table 4: Cluster analysis results and summary

K-means Cluster Method	Metropolitan Areas
Cluster A	Washington, DC; New York, Northern New Jersey, NY-NJ; San Jose-Santa Clara, CA; Chicago, IL; Boston-Cambridge, MA; San Francisco-Oakland, CA; Philadelphia-Camden, PA
Cluster B	Atlanta, GA; Austin, TX; Baltimore, MD; Buffalo-Niagara Falls, NY; Charlotte, NC; Cleveland, OH; Dallas, TX; Denver, CO; Houston, TX; Kansas City, MO-KS; Las Vegas-Paradise, NV; Los Angeles, CA; Memphis, TN; Miami, FL; Minneapolis, MN; New Orleans, LA; Phoenix, AZ; Pittsburgh, PA; Portland, OR; Sacramento, CA; Salt Lake City, UT; San Diego, CA; Seattle-Tacoma, WA; St. Louis, MO-IL; Virginia Beach, VA
Cluster C	Birmingham-Hoover, AL; Cincinnati, OH; Columbus, OH; Detroit, MI; Hartford, CT; Indianapolis, IN; Louisville-Jefferson County, KY-IN; Milwaukee, WI; Nashville, TN; Oklahoma City, OK; Orlando, FL; Providence, RI; Raleigh-Cary, NC; Richmond, VA; Riverside, CA; San Antonio, TX; Tampa, FL; Jacksonville, FL

Figure 2 illustrates where the metropolitan areas belonging to the same cluster groups are geographically located within the entire country. As it is shown in this figure, while cities of all three groups are distributed all around the country, they are not evenly distributed. Figure 1 illustrates the distribution of cluster groups in the four main U.S. regions.[4] In the Midwest and South regions, the highest share belongs to cluster type C, which is the group of sprawled low-density cities. South region also has the same number of cities belonging to the cluster type B and only one city from the type A.

[4] According to the United States Census Bureau

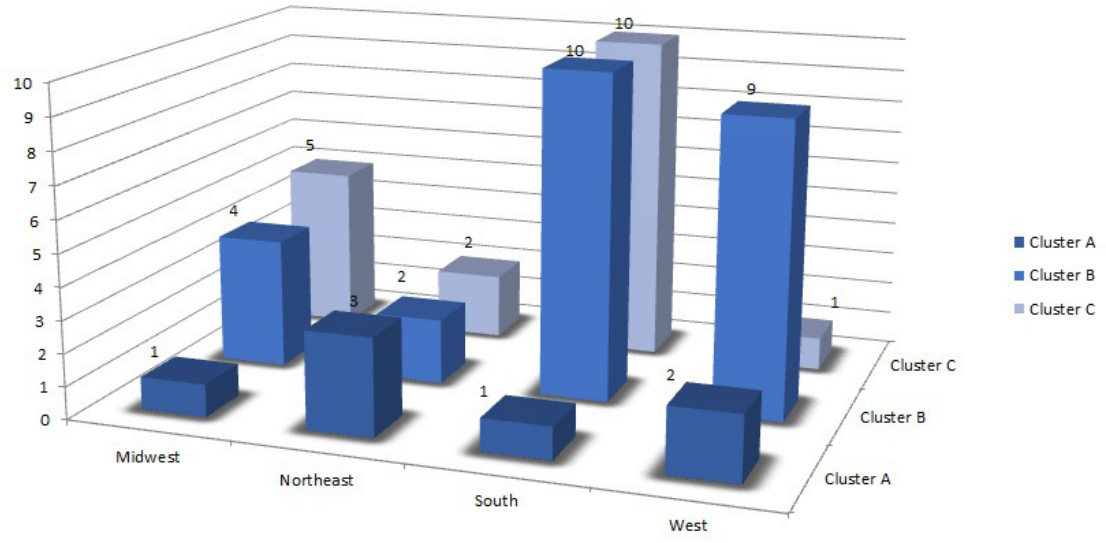

Figure 1: Distribution of cluster groups by U.S. regions

This implies that in the South region, most metropolitan areas follow the medium-to-low density pattern. In the West region, the majority of cases belong to the cluster type B- moderate-density cities with average accessibility and street connectivity-.

There is no generally-accepted rule or measure to evaluate a clustering method based on the output. However, it can be evaluated based on the similarity within and/or dissimilarity between the identified cluster groups. The cluster centroid, a mean of the cluster on each clustering variable, is useful in evaluating the clustering. Interpretation of clustering involves examining the characteristics of each cluster and identifying the similarities/differences. A good method will produce clusters with high intra-group similarity and low inter-group similarity. A method that fails to show substantial variation among the clusters is not recognized as an efficient method and ultimately does not help in understanding the data and find groups in it — as it was the initial goal in clustering the data.

Toward this goal, summary statistics of the main socio-demographic characteristics as well as the land-use measures for each cluster group is provided in Table 5. As it indicates, cluster A shows a higher population and employment densities than the other two groups with a considerable distance. The mean population density in cluster A (21.70) is more than twice as high than that in cluster B (8.64) and about four times as high than that in cluster C (5.85). Similarly, mean employment density in cluster A is more than twice as high than that in cluster B and C. Percentage of population living in residential-only zones is a lot higher in cluster C (0.92 percent) than it is in cluster B (0.76 percent) and A (0.53 percent). The higher this percentage, the lower the accessibility to various destinations and the higher the automobile dependency.

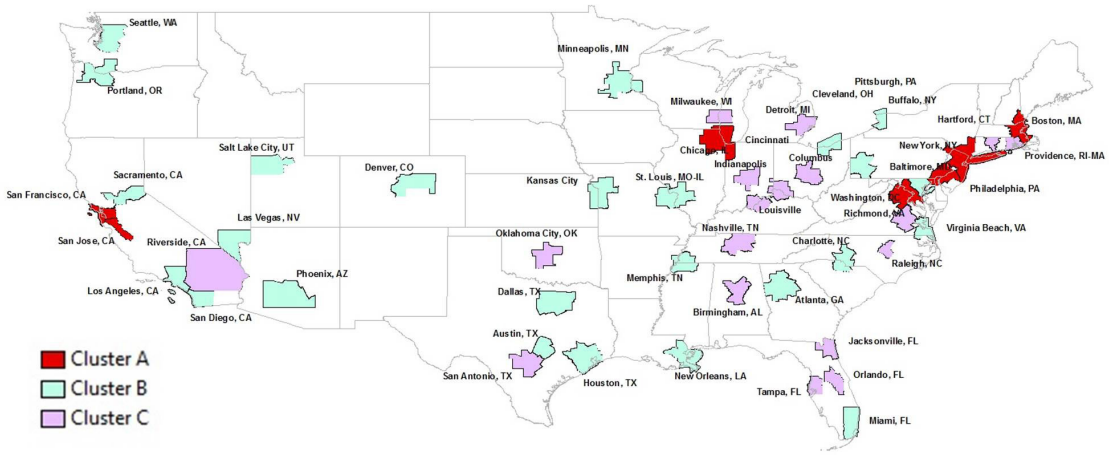

Figure 2: Location distribution of cluster groups

Table 5: Summary statistics by cluster groups

	Cluster A		Cluster B		Cluster C	
	Mean	**Std. dev.**	**Mean**	**Std. dev.**	**Mean**	**Std. dev.**
Socioeconomic and Demographic Characteristics						
Total Employment	3,246,340	2,325,194	1,408,364	1076108	806,993.7	288,623.7
Total Population	7,232,919	5,628,019	3,229,618	2515438	1,925,678	957,611.8
HHs	2,676,953	2,054,958	1,193,101	835,621.4	730,700.3	334,697.6
Workers	2,869,077	2,587,500	1,354,001	1,011,600	776,342.9	336,076.7
Avg_HH_size	2.71	.12	2.65	.17	2.61	.18
P_WrkAge	76.82	1.41	75.35	2.14	75.58	1.84
MedHHInc	86,614.29	15,111.74	69,184	7,733.73	66,622.22	7,482.26
P_AutoOwn0	15.14	8.64	8.66	2.59	7.74	1.78
P_AutoOwn1	35.77	2.81	38.26	4.36	37.97	4.26
P_AutoOwn2+	56.13	8.51	63.44	4.24	65.28	4.23
P_LowWage	22.06	3.17	24.01	2.35	24.81	1.48
P_MedWage	29.21	2.55	35.75	3.07	36.85	3.61
P_HiWage	48.73	5.50	40.24	4.21	38.33	4.39
Built Environment Characteristics						
ResDens_Avg	21.70	13.92	8.64	3.65	5.85	2.11
EmpDens_Avg	7.33	3.96	3.00034	.94	2.18	.64
Entropy_Avg	.46	.050	.46	.048	.48	.065
P_ResOnly	.53	.48	.76	.77	.92	.74
P_LowResDens	36.56	15.83	21.49	14.88	20.83	8.82
P_Hi_ResDens	52.81	20.75	55.86	22.095	50.52	15.19
P_ LowEmpDens	19.13	18.88	5.19	8.28	1.75	2.011
P_ HiEmpDens	35.73	20.69	38.27	21.46	31.54	15.33
E_LowEmpDens	17.29	17.62	3.83	7.28	1.039	2.044
E_HiEmpDens	41.80	19.50	48.78	22.81	46.27	18.59
%SmallBlocks	59.99	5.50	50.91	8.28	48.55	7.76
Block_Size_Avg	.050	.014	.13	.171	.107	.046
Walkscore	74.93	12.81	50.86	14.028	42.13	14.62
Roadnetworkdensity	18.15	32.30	6.026	8.20	16.19	47.29
JobHH_avg	11.29	4.37	12.86	9.58	8.41	10.31
P_D4b050_metro	34.42	7.54	15.49	5.22	1.86	2.57
P_D5br_avg	.65	.42	1.22	2.34	.63	.64
Travel Behavior Characteristics						
VMT*	42,258.72	22,380.78	25,024.99	19,971.65	16,825.69	8,557.73
VMT per capita	6,639.40	1,378.57	7,659.41	853.41	8,787.82	1,024.33
Auto_Commute	75.45	9.45	88.04	3.21	90.99	1.70
Transit_Commute	13.69	8.57	3.49	1.81	1.82	.94
WalkBike_Commute	4.91	1.34	2.80	1.12	2.17	.68
# of observations	7		25		18	

* VMT is measured in million miles

The three cluster groups are somehow similar in terms of average entropy, the percentage of population living in high-residential-density zones and high-employment-density zones and the proportion of employment located in high-employment-density zones. However, the proportion of employment concentrated in low-employment-density zones is a lot higher in cluster A (17.29 percent vs. 3.83 percent and 1.039 percent), which is an indicator of a more evenly distributed pattern for employment. In terms of street connectivity and walkability, cluster A is again in a better shape than the other two groups. The

percentage of small blocks in cluster A is about 60 percent whereas in the other two cluster groups this number is somewhere around 50 percent. Also, the average block size is smaller and the walkscore is larger in cluster A compared to the other two cluster groups.

Looking at transit accessibility measures in Table 5, again it is observed that cluster A has a higher level of transit accessibility than the other two groups (the percentage of metro area's employment located within a ½ mile of transit stations is twice as high in cluster A than in cluster B and compared with cluster C this ratio is about 1/17).

In terms of socio-demographic characteristics, Cluster A has the highest ($86,614) and Cluster C has the lowest ($66,622) median household's income. Similarly, the percentage of high-wage workers in cluster A is higher (48.73 percent) and the percentage of low-wage workers is lower (22.06 percent) compared to the other two cluster groups.

In cluster A there are more households with no automobiles (15.14 percent) and less households with more than 2 automobiles (56.13 percent). The number of households who do not own a car is almost double in cluster A compared to the other cluster groups. This is a very important finding, especially for policy makers who are looking for ways to restrict auto ownership. It is important to note that in areas with a higher accessibility and a more compact, transit-friendly pattern, the percentage of households who decide not to own a car increases, while the median income in these areas is higher and the percentage of high-wage workers is also higher compared to the other groups with a more sprawled and less-connected land-use pattern (see Holtzclaw, Clear, Dittmar, Goldstein, & Haas, 2002; Ewing & Cervero, 2010).

In addition to the car ownership pattern, clusters are also compared based on their VMT and the overall commute mode share pattern. The commute mode share for three modes—auto, transit, and non-motorized transport—has been obtained from the American Community Survey "Journey to Work" data for 2010-2012. The auto mode includes both drive-alone and carpool. Also, worked-at-home population was excluded from the beginning to avoid possible over- and under-estimation of the results. The data was first driven at the county level and then aggregated to the metropolitan level to get the numbers for each of the 50 metropolitan areas.

The total annual VMT for each metropolitan area was calculated using the database provided by the Highway Performance Monitoring System (HPMS) from the Federal Highway Administration. This dataset provides traffic volume data by road segments for all road types. Similar to the mode share calculation, the annual VMT for the year 2008 (the most updated data available) was first calculated for each county and then aggregated to the whole metropolitan area to get the actual VMT number. The following formula shows how the annual VMT at the metropolitan level was calculated:

$$VMT_i = \sum_{j=1}^{n} \sum_{k=1}^{n} AADT_k L_k * 365 \qquad (2)$$

where:

VMT_i = Annual VMT for the metropolitan area i, $1 \le i \le 50$

$AADT_k$ = Average annual daily traffic for the road segment k

L_k = Segment length, mile

j is the county's identifier and ranges from one to the number of counties in metropolitan area *i*.

The analysis shows that cluster A has a higher overall VMT compared to the other clusters but lower per capita VMT. The share of auto commute in cluster A is about 75 percent while in cluster B it is 88 percent and in cluster C it is about 91 percent. Similarly, the share of transit commute in cluster A is about 14 percent which is about three times higher than that in cluster B and about 7 times higher than that in cluster C. The same pattern exists for the share of walk/bike mode for commuting trips.

4 Conclusions and future research

This paper seeks to contribute to a better understanding and quantification of the overall physical form of metropolitan areas by proposing a range of multi-level multi-dimensional measures to explore the urban structure at higher geographical levels. If we do not understand and empirically explore various dimensions and characteristics of metropolitan-level built environment, then the policies proposed to cope with congestion and improve transit ridership through the entire metropolitan areas, such as smart growth and transit-oriented development would not be as effective. Thus, these unique measures could serve as a foundation on the debate about the relationship between travel behavior and the built environment at both small and large scales.

The proposed measures of metropolitan-level built environment are calculated for 50 metropolitan areas across the country. These measures can facilitate research on the relationships among land use and urban form and various dimensions of urban daily travel behavior by providing a clearer and more comprehensive picture of the overall structure of urban areas measured at various hierarchical levels.

This comprehensive quantification of urban form allows for a better understanding and visualization of various aspects of urban form, which could potentially be used in various analyses of the relationship between land use and transportation, environment, housing market, and more. It also facilitates planning and evaluation of various land-use policy scenarios. The proposed measures and indices have been calculated for the 50 most populous urban areas in the U.S., and the consistency of the values allows for several comparative analyses in these metropolitan areas.

Thus, it can be again implied that cities with a more compact land-use pattern, have an overall lower automobile dependency and higher level of transit and non-motorized mode share, similar to what several previous studies have found. The higher overall VMT in cluster A is a direct result of the size of the metro areas in this group and higher population living in these cities.

Overall, these findings are potentially significant to future land use and transportation planning projects, and they could be used to better utilize land use and transportation planning and policy analyses used by planners and researchers. Clustering of urban areas would eventually help policy- and decision makers in their decision-making process to examine and/or evaluate new and old land use and transportation policies and planning scenarios, identify similar patterns, and understand how similar policies implemented in urban areas with a similar urban form structure would result in a more efficient and successful planning for the future.

In the next steps, measures developed in this study can be used to group the metropolitan areas based on their centrality and employment distribution pattern into monocentric, polycentric, decentralized, and sprawled patterns. This will also help researchers investigate the effects of urban form and development pattern on many topics, such as the housing and real estate market, transportation investments, and economic development.

Acknowledgements

This research was partially supported by the National Transportation Center at the University of Maryland, College Park, which is one of the five MAP-21 National University Transportation Research Centers funded by the U.S. Department of Transportation. The authors are solely responsible for all statements in the paper.

References

Bento, A., Cropper, M., Mobarak, A., & Vinha, K. (2005). The effects of urban spatial structure on travel demand in the United States. *The Review of Economics and Statistics, 87*(3), 466–478.

Borgen, F. H., & Barnett, D. C. (1987). Applying cluster analysis in counseling psychology research. *Journal of Counseling Psychology, 34*(4), 456.

Cervero, R., & Murakami, J. Effects of built environments on vehicle miles traveled: Evidence from 370 US urbanized areas. Environment and Planning A 42(2), 400–418.

Ewing, R. (1994). Characteristics, causes, and effects of sprawl: A literature review. *Environmental and Urban Issues, 21*(2), 1–15.

Ewing, R. (1997). Is Los Angeles-style sprawl desirable? *Journal of the American Planning Association, 63*(1), 107–126.

Ewing, R., & Cervero, R. (2010). Travel and the built environment: A meta-analysis. *Journal of the American Planning Association, 76*(3), 265–294.

Fischer, M. M., & Getis, A. (Eds.). (2009). *Handbook of applied spatial analysis: software tools, methods and applications.* Berlin: Springer Science & Business Media.

Frenkel, A., & Ashkenazi, M. (2008). The integrated sprawl index: Measuring the urban landscape in Israel. *The Annals of Regional Science, 42*(1), 99–121.

Galster, G., Hanson, R., Ratcliffe, M. R., Wolman, H., Coleman, S., & Freihage, J. (2001). Wrestling sprawl to the ground: Defining and measuring an elusive concept. *Housing Policy Debate, 12*(4), 681–717.

Gordon P., Richardson, H., & Jun, M.-J. (1991). The commuting paradox: Evidence from the top twenty. *Journal of the American Planning Association, 57*(4), 416–420.

Gordon, P., Kumar, A., & Richardson, H. W. (1989). Congestion, changing metropolitan structure, and city size in the United States. *International Regional Science Review 12*(1), 45–56.

Grengs, J., Levine, J., Shen, W., & Shen, Q. (2010). Intermetropolitan comparison of transportation accessibility: Sorting out mobility and proximity in San Francisco and Washington, D.C. *Journal of Planning Education and Research, 29*(4), 427–443.

Holtzclaw, J., Clear, R., Dittmar, H., Goldstein, D., & Haas, P. (2002). Location efficiency: Neighborhood and socio-economic characteristics determine auto ownership and use-studies in Chicago, Los Angeles and San Francisco. *Transportation Planning and Technology, 25*(1), 1–27.

Ingram, G. (1998). Patterns of metropolitan development: What have we learned? *Urban Studies, 35*(7), 1019–1035.

Kaufman, L. R., & Rousseeuw, P. J. (1990) *Finding groups in data: An introduction to cluster analysis.* Hoboken, NJ: John Wiley & Sons Inc.

Kelly-Schwartz, A. C., Stockard, J., Doyle, S., & Schlossberg, M. (2004). Is sprawl unhealthy? A multilevel analysis of the relationship of metropolitan sprawl to the health of individuals. *Journal of Planning Education and Research, 24*(2), 184–196.

Malpezzi, S., & Guo, W. K. (2001). *Measuring sprawl: Alternative measures of urban form in US metropolitan areas*. Unpublished manuscript, Center for Urban Land Economics Research, University of Wisconsin, Madison.

McCann, B. A., & Ewing, R. (2003) Measuring the health effects of sprawl: A national analysis of physical activity, obesity and chronic disease. Smart Growth America Surface Transportation Policy Project. Retrieved from http://www.smartgrowthamerica.org/healthreportpr.html

Nasri, A., & Zhang, L. (2012). Impact of metropolitan-level built environment on travel behavior. *Transportation Research Record, 2323*(1), 75–79.

Nasri, A., & Zhang, L. (2014). Assessing the impact of metropolitan-level, county-level, and local-level built environment on travel behavior: Evidence from 19 US urban areas. *Journal of Urban Planning and Development, 141*(3), 04014031

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, 20*(2), 134–148.

Schwanen, T., Dieleman, F. M., & Dijst, M. (2004). The impact of metropolitan structure on commute behavior in the Netherlands: A multilevel approach. *Growth and Change, 35*(3), 304–333.

Smith, J., & Saito, M. (2001). Creating land-use scenarios by cluster analysis for regional land use and transportation sketch planning. *Journal of Transportation and Statistics, 4*(1), 39–49.

Torrens, P. M., & Alberti, M. (2000). Measuring sprawl. Working Papers 27. Centre for Advanced Spatial Analysis, University College London. Retrieved from http://www.casa.ucl.ac.uk

Tsai, Y. H. (2005). Quantifying urban form: Compactness versus sprawl. *Urban Studies, 42*(1), 141–161.

Veneri, P. (2010). Urban polycentricity and the costs of commuting: Evidence from Italian metropolitan areas. *Growth and Change, 41*(3), 403–429.

Yang, J., French, S., Holt, J., & Zhang, X. (2012). Measuring the structure of US metropolitan areas, 1970–2000: Spatial statistical metrics and an application to commuting behavior. *Journal of the American Planning Association, 78*(2), 197–209.